



**HAL**  
open science

# Optimized Assessment of Physical Rehabilitation Exercises using Spatiotemporal, Sequential Graph-Convolutional Networks

Ikram Kourbane, Panagiotis Papadakis, Mihai Andries

► **To cite this version:**

Ikram Kourbane, Panagiotis Papadakis, Mihai Andries. Optimized Assessment of Physical Rehabilitation Exercises using Spatiotemporal, Sequential Graph-Convolutional Networks. 2024. hal-04117417v2

**HAL Id: hal-04117417**

**<https://imt-atlantique.hal.science/hal-04117417v2>**

Preprint submitted on 16 Dec 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

# Optimized Assessment of Physical Rehabilitation Exercises using Spatiotemporal, Sequential Graph-Convolutional Networks

Ikram Kourbane\*, Panagiotis Papadakis and Mihai Andries

IMT Atlantique, Lab-STICC, UMR CNRS 6285, team RAMBO, F-29238 Brest, France,

## ARTICLE INFO

**Keywords:**  
Rehabilitation  
Quality score assessment  
Classification  
Graph convolutional networks  
Spatiotemporal

## ABSTRACT

Rehabilitation is the process of helping people regain or improve lost or impaired function due to injury, illness, or disease. To assist in tracking the progress of patients undergoing rehabilitation, this paper proposes a lightweight graph-based deep-learning model for the automatic assessment of physical rehabilitation exercises. The model takes as input the 3D skeleton sequence of a patient performing a movement and outputs a continuous quality score, as a means for patient supervision that could complement or even substitute the need for ordinary clinical exams. Two graph convolutional networks (GCNs) are sequentially employed to learn spatial and temporal features, the first learning key joint relationships per exercise category and the second exploiting frame correlation to focus on relevant parts of the input sequence. Furthermore, in order to enhance the significance of the scores derived from testing, we propose implementing a classification phase. This phase enables the regression model to produce scores exclusively for sequences specifically tailored to each exercise, which further ensures that an input sequence is assessed only if it corresponds to a complete movement demonstration. The evaluation of the proposed approach on the publicly available KIMORE and UI-PRMD datasets shows that our approach outperforms the state-of-the-art in terms of quality score prediction as well as in terms of efficiency. Our project page is available online.

## 1. Introduction

Human motion analysis is a highly active research area in computer vision. While most studies in the field focus on action detection and recognition [1, 2, 3, 4], few address the human movement quality assessment (HMQA) task, which identifies and quantifies possible deviations from valid movement patterns and optionally provides feedback on the manner in which a person performs an action. HMQA has applications in several domains, including functional capacity evaluation, sports movement optimization, ergonomic risk assessment, physical therapy and rehabilitation settings. According to estimates from the World Health Organization, approximately 2.4 billion people need physical rehabilitation treatments to recover from surgeries or manage various musculoskeletal disorders [5]. This number is constantly increasing as the prevalence of chronic diseases and injuries rises.


On the other hand, assessment of human movement in order to quantify the level of physical impairment requires extensively trained physiotherapists and doctors. However, human evaluation is bound to several limitations, including the maximum permitted of patients per doctor, the long duration and high cost of the examination procedure, and the presence of human-evaluator bias. Additionally, the lockdown during the COVID-19 pandemic further increased the demand for a safe home-based rehabilitation system that uses ordinary sensors to obtain skeletal data. This mitigates the need for using high-edge and expensive Motion Capture (MOCAP) systems, whose dependence on wearable markers and multiple distributed sensors that need to be carefully calibrated may be overly restrictive or cumbersome. From

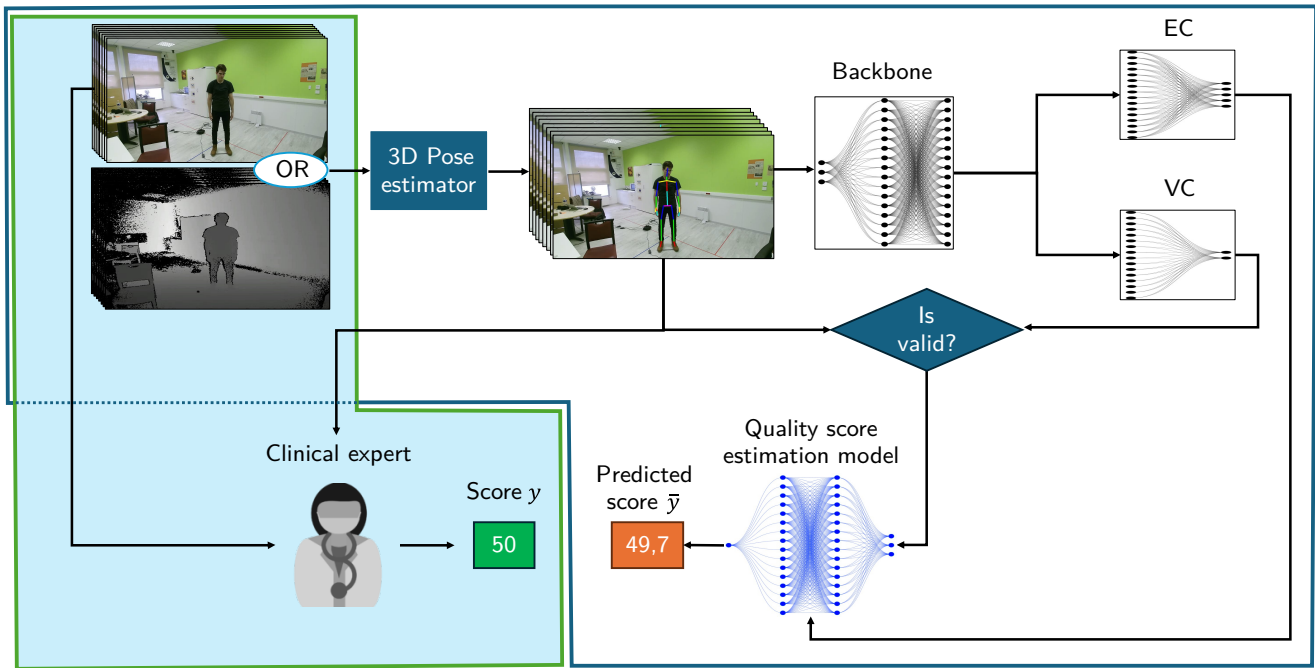
the patient's point of view, MOCAP systems further incur a sense of discomfort that could result in unnatural movements and in turn, less reliable assessments. Currently, MOCAP systems are only available in clinical settings, which unavoidably limits access to rehabilitation services.

In this regard, computer vision-based methods present an effective and economical solution in remote areas. They employ ordinary RGB cameras or low-cost RGB+Depth sensors that are readily available and more affordable while they do not require physical contact with the subject [6]. To predict the quality score of the input human skeleton trajectory, they typically employ deep learning techniques, such as convolutional neural networks (CNN) [7], often in conjunction with long short-term memory networks (LSTM) [8], so as to extract spatial and temporal features [9]. However, these methods do not explicitly exploit the topological structure information of the human body. To address this limitation, seminal studies have shown the potential of graph neural networks (GNNs) [10, 11] for motion quality assessment. However, these methods use fixed hand-crafted joint relationships (adjacency matrix) to describe the connections between human body joints. This can be restrictive, as it can be difficult to capture the dependencies between physically disconnected parts of the body. Furthermore, fixing the connections between joints for all actions may limit spatial learning since each action should have different joint relationship constraints.

Another limitation of existing HMQA methods is that they opt for training a separate model for each exercise, which is computationally expensive and requires a lot of training data, especially when there are many exercises to be considered [12]. These observations instigated us to train a single model to assess all physical rehabilitation exercises, rendering our approach more efficient to train and deploy.

\*Corresponding author

 ikram.kourbane@imt-atlantique.fr (I. Kourbane)



**Fig. 1:** The proposed pipeline analyzes 3D joint positions extracted from depth or RGB data through two primary components: the exercise classifier (**EC**) and the validity classifier (**VC**). The validity classifier assesses the coherence of the sequence before the quality score estimation (blue polygon). Our quality estimation model is trained to match the score of the clinical expert who relies on a video stream as well as the 3D pose sequence so as to provide the ground truth score (green polygon).

Our network consists of two main modules, namely a spatial GCN and a temporal GCN, which are trained in an end-to-end manner. The spatial GCN learns an adjacency matrix for each exercise to capture the implicit connections and important spatial information between the joints. The temporal GCN takes the output of the spatial GCN as input to model the temporal evolution of the joint movements. It considers the frames of the motion sequence as nodes in a graph and links them using a learnable Gaussian-like adjacency matrix. This allows the model to learn temporal attributes by aggregating information from temporally close frames.

Furthermore, we observe that current HMQA methods rely on the very strong assumption of perfect exercise selection and execution [12, 13, 14, 9, 11]. However, real-world scenarios frequently encompass variations in performance, leading to inaccurate quality scores when users deviate from the instructed exercise execution. To address this challenge, our proposed approach incorporates two classifiers (Fig 1). Firstly, an exercise classifier identifies the action category before applying the regression. This ensures that the quality score estimation model receives exercise-specific data. The secondary classifier serves to validate the input sequence by discerning between valid and invalid exercise sequences. This prioritization ensures that scoring focuses on entire movements rather than isolated segments, effectively capturing the dynamic essence of real-world exercise and enhancing score precision across the entire session. Furthermore, this classifier acts as a safeguard, preventing the model

from being misled by irrelevant or out-of-distribution movements. Since the existing rehabilitation datasets are small-scale [15, 16], we conduct various data augmentations and preprocessing techniques to improve the performance of our model and avoid overfitting. In summary, the contributions of this work are as follows:

- We introduce an end-to-end lightweight HMQA network that adapts GCNs to learn per-exercise adjacency matrices to capture the most relevant joint connections. In addition, it extracts temporal features using the correlation between nearby frames.
- We incorporate a preceding action classification step to prompt the model to generate quality scores exclusively for valid sequences. This approach ensures that the generated scores are pertinent to the context of the given sequences.
- We conducted extensive experiments on two rehabilitation datasets (KIMORE [15] and UI-PRMD [16]) achieving superior performance compared to the state-of-the-art. Additionally, our method is computationally efficient and scalable, making it suitable for real-time applications.

The rest of this paper is organized as follows. Section 2 reviews the related studies of our work. Section 3 explains the proposed method in depth and describes the most important modules of our framework. Section 4 describes the experimental settings. Section 5 analyzes the obtained results

on two public datasets and compares the proposed approach against the state-of-the-art. Finally, Section 6 presents the conclusion of the study and the direction of future work.

## 2. Related work

In this section, we review the related assessment of physical rehabilitation exercises (APRE) methodologies, alongside a discussion of associated literature in the domain of GCNs that aligns with the scope of our research.

### 2.1. Assessment of physical rehabilitation exercises

In recent years, there has been a growing interest in using artificial intelligence-based techniques to improve the accuracy and efficiency of APRE [17, 18]. Early studies focused on probabilistic approaches like Hidden Markov models (HMMs) [19, 20] and mixtures of Gaussian distributions [21] for assessing exercises. However, these approaches require several preprocessing stages, such as feature extraction and data cleaning. This can be time-consuming, and computationally expensive while identifying the optimal parameter values that lead to the best performance can be particularly challenging.

End-to-end deep learning models have demonstrated the capacity to automatically assess a patient's physical abilities and limitations based on data collected from wearable [22, 23] or vision sensors [24, 25]. First-generation methods typically classify movements as either correct or incorrect [26], without however providing details on the quality of the movement. More recent methods overcome this limitation by predicting a continuous score for each movement [9, 13, 27, 28] that can be more informative and allow the monitoring of subtle progress over time.

Du et al. [29] introduced a method to quantify patient performance using a Gaussian Mixture Model (GMM) log-likelihood metric. They employed a scoring function to map these performance metrics to a movement quality score within the range of 0 to 1. Their model utilized hierarchical processing of joint displacements across various body parts, incorporating convolutional and recurrent layers to encode correlations in movement data. Also, Kanade et al. [30] proposed a transformer-based architecture and showed that using data augmentations for generating movement quality scores results in significant performance boosts over existing methods. However, these methods do not explicitly consider the topological structure of the human body. This means that they do not take into account how the different parts of the body are connected to each other and how they move together. To address this limitation, many recent approaches use GNNs to model skeletal constraints among neighboring joints in a non-Euclidean space.

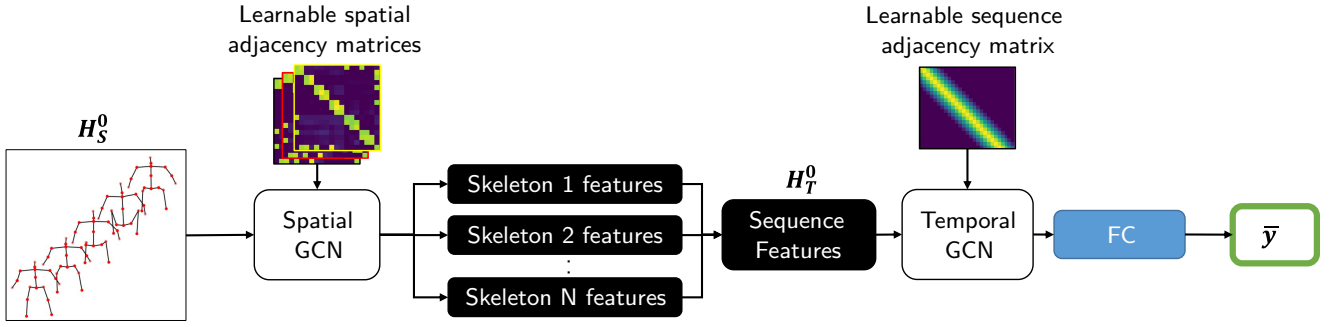
### 2.2. GCNs for action recognition and APRE

GNNs can extract features from data that are arranged in an irregular graph structure. More particularly, GCNs have been effectively applied to various tasks that involve analyzing skeleton sequences, such as in the domain of gesture classification [31, 32], action recognition [33, 34, 35, 36],

and HMQA [37, 38, 39]. The spatio-temporal graph convolutional networks (ST-GCN) framework proposed by Yan et al. [11] is the seminal work that captures both spatial and temporal features from skeleton data, achieving remarkable results in classifying actions.

In subsequent works, Chowdhury et al. [40] proposed a model that uses a GCN to extract features from skeleton data, followed by an LSTM to predict the output quality score of an exercise. Chen et al. [41] proposed an ensemble-based graph convolutional network (EGCN) for movement assessment, which uses a combination of multiple GCNs to learn more robust features from the movement data. Deb et al. [12] proposed a GCN-based method that can process variable-length inputs using LSTMs and employs self-attention of body joints indicating their role in predicting assessment scores. Following this work, [42] merges modified STGCN and transformer architectures to handle spatio-temporal data effectively and identify the most important joints. The attention mechanism within the transformer encoder selectively focuses on pertinent segments of the input sequence. Also, Réby et al. [24] used a transformer network to learn the long-range dependencies in the input data, using a graph network to learn the spatial and temporal relationships between the different body joints. However, these methods do not identify the main joint connections for each distinct exercise category, which could improve training and testing performances. Additionally, they are computationally demanding, either because a separate model must be trained for each exercise [12, 41] or because they use other complex techniques such as LSTMs or transformers [40, 11]. In contrast, our approach trains a single model that learns the most important joint and frame connections based solely on GCNs, which increases the efficiency of the method.

Some other related works in the action recognition field revolve around learning a per-action adjacency matrix. More particularly, [43] leverages ST-GCN in a multi-task learning framework to dynamically adjust the adjacency matrix, enhancing its ability to represent intricate patterns and relationships between joints. [44] introduces an adaptive adjacency matrix designed with a unique partitioning strategy for neighbor sets. This strategy decomposes the adjacency matrix into three parametric matrices, enabling more flexible and efficient feature extraction. Also, [45] integrates an adjacency matrix generation module, which pre-analyzes node sets and generates an adaptive adjacency matrix tailored to the input data characteristics. Unlike previous methods that utilize spatial and temporal layer blocks, our approach prioritizes the identification of spatial relationships among joints within individual skeletons. We then focus on learning temporal features between skeletons in sequence. This methodology offers advantages in capturing dependencies among joints and frames, as demonstrated in Section 5.



**Fig. 2:** The overall architecture of the proposed end-to-end GCN-based method (cf. *Quality score estimation model* in Figure 1) is as follows: The input sequence skeleton data  $H_S^0$  is fed to the Spatial GCN to learn the spatial structure of each frame, corresponding to the class of the input exercise to be optimized. The learnable features  $H_T^0$  of the spatial GCN are then concatenated and fed to the temporal GCN, which extracts features for the entire sequence based on a learnable adjacency matrix. Finally, the output of the temporal graph is fed to a fully connected (FC) layer to estimate the quality score  $\bar{y}$  of the input exercise sequence.

### 3. Methodology

Human actions can be viewed as a set of spatio-temporal changes in motion. Inspired by the natural graph representation of the human body, we use GCNs (Section 3.1) to learn the relationships between the joints in the human body. Fig. 2 shows the architecture of our method, which consists of spatial and temporal GCNs trained in an end-to-end manner (Section 3.2 and Section 3.3). We apply a classification stage (Section 3.4) before the quality score regression to ensure that the input sequence is permissible for training or testing. The overall pipeline is complemented with different data augmentation and pre-processing techniques that are detailed in Section 4.2.

Unlike current GCNs-based action recognition methods that jointly interweave spatial and temporal processing [43, 11], our sequential approach decomposes spatial and temporal analysis to gain a deeper understanding of both aspects. We begin by meticulously characterizing the spatial relationships between joints within each individual frame. This comprehensive analysis allows us to capture the intricate interactions and dependencies within each skeletal pose. Subsequently, we leverage this fine-grained understanding of spatial configurations to learn how these relationships evolve over time. As Fig. 2 shows, we concatenate the extracted features from all frames to serve as input to the temporal GCN that also learns the best frame connection pattern.

The output of the Temporal GCN is then directed to a set of linear layers to estimate the quality score of the input sequence. Given that we are dealing with a regression problem, we employ the  $L1$  loss function to quantify the difference between the predicted value and the ground truth quality score:

$$L1 = \sum_{i=1}^s (\|y - \bar{y}\|) \quad (1)$$

where  $s$  is the number of sequences in the dataset and  $y$  and  $\bar{y}$  are the ground-truth and predicted quality score values, respectively.

#### 3.1. Graph Convolutional Networks

GCNs are a type of neural network that can be used to learn representations of nodes in graphs. A graph, denoted by  $G = (V, E)$  is a data structure that consists of a set of nodes,  $V$ , and a set of edges,  $E$ , where the edges represent connections between the nodes. GCNs apply convolution operations by taking into account the relationships between the nodes in a graph represented by the adjacency matrix  $A$ , enabling them to learn more complex patterns than ordinary neural networks.

Specifically, an adjacency matrix  $A$  is a square matrix used to represent a graph. For a graph with  $V$  nodes, the adjacency matrix  $A$  is a  $V \times V$  matrix where each element  $A_{i,j}$  indicates the presence or absence of an edge between nodes  $i$  and  $j$ . The values in the matrix can be binary or weighted.

In human motion analysis, the adjacency matrix encodes crucial information about the relationships between nodes. It helps the model understand which nodes (e.g., body joints) are more closely related, either spatially (e.g., anatomically close joints) or temporally (e.g., movements occurring in close succession). This structure enables the GCN to effectively aggregate and propagate information across the network.

In our pipeline, the adjacency matrix is initially constructed based on a predefined human-topology. During training, this matrix is updated to better capture the relationships between nodes, adapting to each exercise specific characteristics. The reconstructed matrix converges towards a form that emphasizes connections based on both spatial and temporal proximity.

The adjacency matrix is used in the convolutional operations of the GCN to determine how features from different nodes are combined. The matrix guides the convolution process, ensuring that relevant information is shared between connected nodes, thereby improving the model's ability to learn meaningful patterns in the data. The propagation rule typically used in GCNs is defined as:

$$H^{k+1} = \text{ReLU}(AH^k W^k) \quad (2)$$

where  $ReLU$  is the activation function,  $H^k$  and  $W^k$  are the 3D positions and the weights in the  $k^{th}$  layer, respectively, where  $k \geq 0$ . Our two classifiers, along with the spatial and temporal GCNs, consist of two layers each (which sets  $k=2$ ), that have the internal structure presented in [10]. The number of frames in the temporal GCN is fixed to a constant  $M$ .

### 3.2. Learning spatial features using per-action adjacency matrix

We learn a separate adjacency matrix for each of  $n$  different types of exercises so as to allow variable importance to the body joint connections (Fig. 2). During inference, our exercise classifier **EC** selects the adjacency matrix that corresponds to the input sequence. To allow variable joint connectivity depending on the temporal and spatial context while assessing rehabilitation exercises, we generalize the basic propagation rule to the following equation:

$$H_S^{k+1} = ReLU\left(\mathbf{EC}((A_1|A_2|\dots|A_n)H_S^k W_S s^k)\right) \quad (3)$$

where  $k \geq 0$  and the  $H_S^0$  represents the input feature matrix for the first layer. Each row corresponds to the 3D positional coordinates of a node (joint) in the input skeleton sequence. We obtain the features for the second layer by multiplying the adjacency matrix with  $H_S^0$  and the learnable weights. The adjacency matrix is selected by **EC** and is optimized to identify the nodes that will contribute to information aggregation for that specific movement class.

Such an approach is advantageous compared to a conventional setting of the adjacency matrix  $A$  according to a predefined human body topology [11], that limits the ability of the model to capture the feature information of nodes that are far apart in the topology of the human body graph. For example, the "squatting" action (cf. Figure 1 in [15]) requires the model to be able to capture the relationship between the shoulder and knee joints, despite being physically disconnected in the topology of the human body graph. As a result, the model may not be able to learn to capture the subtle nuances of this action because of the absent connections. Another example can be observed in the context of actions involving intricate movements spanning multiple body parts. For instance, actions like pelvis rotation demand a model capable of understanding the interplay between the pelvis, spine, and shoulder joints. A global adjacency matrix might miss these crucial connections, impairing the ability of the model to accurately recognize and assess such complex movements. Hence, incorporating a more flexible adjacency matrix that captures a greater variability of movements can significantly enhance performance and generalizability.

Furthermore, fixing the adjacency matrix in all graph convolutional layers and input samples is not the best choice for action analysis, because the relevant connections between joints can vary depending on the action being performed. In other words, different human actions may not necessarily share the same relationship constraints, since they do not have the same joint configurations. For example,

when doing a squat, the main joints involved are the knees, hips, and ankles. However, when doing a push-up, the main joints involved are the shoulders, elbows, and wrists.

Therefore, we suggest learning  $n$  adjacency matrices, denoted as  $A_i$ , corresponding to the  $n$  in total possible movements. These matrices are updated within the same model which allows us to identify the relevant joint connections during training. The adjacency matrices are initialized with the skeleton-based topology, where each joint is only connected to its immediate neighbors. The parameters of these matrices are optimized during training using the  $L1$  loss. In test time, we learn the adjacency matrix corresponding to the input skeleton sequence class identified by our action classifier **EC**.

### 3.3. Learning temporal features using GCNs

Previous APRE methods employ GCNs to extract spatial features and incorporate an LSTM [8] or transformers [46] at the end of the architecture to extract temporal features from the sequence.

Our approach extends GCNs to learn temporal information directly, without the need for an additional temporal module (Fig. 2). We achieve this by considering the frames of the motion sequence as nodes in a graph and linking them using a learnable temporal adjacency matrix  $A^G$ . We have observed that the learnable frame connections converged to a Gaussian-like adjacency matrix. This indicates that the model is more attentive to information from temporally close frames than temporally distant frames, which is essential for understanding the dynamic nature of human motion. For example, the model can learn that the frame in which a person's knee begins to bend is temporally close to the frame in which their hip begins to bend when performing a squat. The forward pass of the temporal GCN is defined as follows:

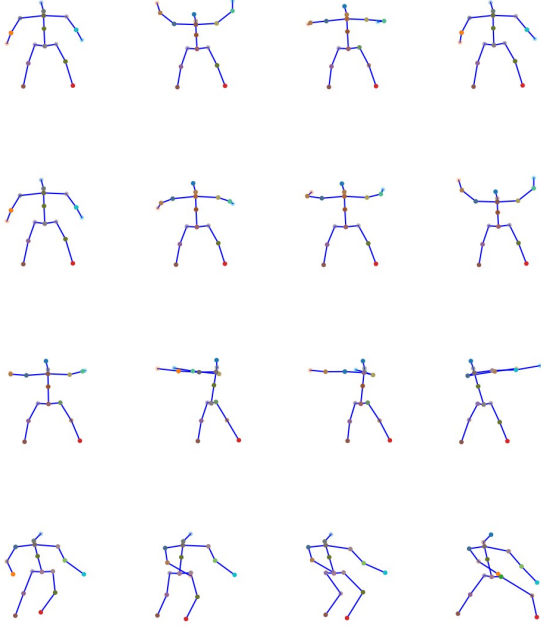
$$H_T^{k+1} = ReLU(A^G H_T^k W_T^k) \quad (4)$$

where  $k \geq 0$  and  $H_T^0$  are the output learnable spatial features of the spatial GCN, which serve as the input features for the first layer in our temporal GCN.

This approach proves more effective than previous GCN-based temporal modules, such as [11], which restrict learning to only the previous and next frames. This is because our method allows the model to learn temporal dependencies over a broader range of frames, capturing more nuanced motion patterns and transitions. Consequently, the model gains a richer and more comprehensive understanding of the temporal dynamics in human motion sequences, leading to improved performances (see Section 5)

### 3.4. Out-of-distribution sequence identification

Current APRE approaches often operate under the unrealistic assumption that only valid sequences will be present during testing, thus generating scores regardless of the input sequence's validity. In other words, they presuppose that users will execute exercises in accordance with the prescribed guidelines and without any interruptions in movement sequences. Additionally, users are typically expected



**Fig. 3:** Examples of possible received sequences during testing time. Rows from top to bottom represent: a valid lifting arm movement, an incomplete lifting arm movement, an incorrect exercise class represented by a trunk rotation movement from the KIMORE dataset [15], and a dancing movement sequence from an out-of-distribution dataset [47].

to designate which exercise class is about to be executed. However, this assumption overlooks the inherent likelihood of human errors and inconsistencies during exercise execution. Such errors can significantly impact the accuracy of quality scores generated by these models, rendering them unreliable for providing meaningful feedback. Therefore, it is imperative for the model to possess the capability to identify the situations, which are illustrated in Figure 3:

1. Incompletely executed movements
2. Incorrect exercise classes
3. Out-of-distribution movements

Our approach addresses the limitations of previous methods that struggled with scoring such sequences, by using the previously described **EC** as well as a validity classifier **VC** (cf. Figure 1). By incorporating an exercise classification stage before regression, we enforce that the model will generate a score for the correct exercise class. To do so, we train our classifier using a dataset that represents in-distribution data [15]. By incorporating this capability, users can receive reliable feedback while our **EC** can select the adjacency matrix that should be used in test time.

Sequences may sometimes be incomplete or include movements that are not part of the dataset. These out-of-distribution movements are either underrepresented or entirely absent in the training data used to develop the model. Since the model is unfamiliar with these sequences,

accurately estimating the quality score becomes challenging. Additionally, clinicians may not be interested in these movements as they do not correspond to the typical or expected patterns relevant to the specific clinical context or diagnosis. Therefore, identifying and filtering out these out-of-distribution movements is crucial to ensure that the model outputs are clinically meaningful and reliable.

The role of the **VC** is to output a probability indicating the validity of the sequence. To establish the ground truth for training our **VC**, we manually remove outliers that are caused by sensor errors or subjects engaging in activities unrelated to the intended exercise, such as talking or walking at the beginning of the exercise. Keeping only valid movements makes the data more representative of the underlying distribution.

After that, we randomly select a percentage between 25% and 75% of the length of the original sequence to determine the length of the incomplete sequence. In the sequel, another random number is chosen to designate the starting position of cropping. The newly extracted sub-sequence is then annotated as class 1, while the original complete sequence is annotated as class 0. Additionally, we include sequences from the NTU-60 dataset [47] into class 1 to represent out-of-distribution movements. This augmentation further enhances the model's ability to discern between relevant and irrelevant movements, enhancing its robustness and adaptability to diverse scenarios.

We perform down/up sampling to standardize the size of all cropped sequences and NTU sequences to a fixed constant  $M$ . We employ the *CrossEntropy* loss function to classify both the exercise class and the validity of the input sequence as expressed as follows:

$$L_{CE} = \sum_{i=1}^c t_i \log(p_i) \quad (5)$$

where  $n$  is the number of classes,  $t_i$  is the ground truth label and  $p_i$  is the *SoftMax* probability for the  $i^{th}$  class.

## 4. Experimental settings

In this section, we thoroughly detail the datasets used, including their sources and characteristics, while also recalling the evaluation metrics for model performance. We further detail the protocols used for training and evaluation for the sake of clarity and result reproducibility. Finally, we explore data augmentation techniques aimed at bolstering model robustness and generalization, and we elaborate on preprocessing methods employed to prepare the data for optimal model performance.

### 4.1. Datasets

We conducted our experiments using two publicly available, rehabilitation exercise datasets.

The **KIMORE** dataset [15] is a valuable resource for research on human motion analysis and rehabilitation. It is a well-curated dataset that has been carefully annotated by medical experts. The KIMORE dataset includes a variety

of low-back pain exercises and has three data inputs: RGB, depth videos, and skeleton positions for 25 joints acquired using a Kinect sensor. It was collected from 78 subjects, including 44 healthy subjects and 34 patients with pain and postural disorder (Parkinson, back-pain, stroke). This dataset also provides a set of clinical features, which are invariant among people and selected on the basis of the scope of the exercise.

The **UI-PRMD** dataset [16] contains human motion data collected from healthy individuals performing ten common rehabilitation exercises targeting different body regions. The dataset includes positions and angles of the body joints in the skeletal models provided by the Vicon and Kinect sensors. For each exercise, ten healthy subjects perform ten repetitions in both a correct and incorrect manner. Each sequence is about 20 seconds, and the number of joints is 25 and 39 for Kinect and Vicon, respectively. The performance scores are generated based on a Gaussian mixture model. A scoring function is defined to map the performance metric values into movement quality scores in the range [0, 1]. Since this dataset is collected from healthy individuals, the data may be less representative of the movements of patients with injuries or disabilities.

## 4.2. Data augmentation and pre-processing

*Data augmentation* Due to the scarcity of annotated data, there is a lack of rehabilitation exercise datasets. The KIMORE [15] and UI-PRMD datasets [16] are small-scale and suffer from a data imbalance problem, where healthy people outnumber unhealthy people by a large margin. In this respect, training a model without data augmentation could be problematic and lead to overfitting.

To alleviate this problem, we augment the size and diversity of the datasets by generating new motion sequences from the existing data. To generate sequences of different speeds, we randomly add or remove  $L$  frames, respectively. To ensure that the quality score of the newly generated sequence is still relevant to the original, the selected random number  $L$  is in the range [0%,25%] of the sequence length.

Relying on feedback received from clinicians, adjusting the speed of the original sequence does not compromise the quality score. This is because individuals may perform actions at varying speeds, influenced by factors such as age and physical condition. Besides, we conducted several experiments to empirically validate this configuration for data augmentation. We note that the used data augmentation does not compromise the sequence's validity, as the added or removed frames are not consecutive. In contrast, the sequence becomes invalid for the classifier if we crop or add blocks of consecutive frames omitting significant motion segments (25% and 75% of the length of the sequence), which makes the linear interpolation harder.

Lastly, to enhance the dataset's diversity and robustness, we employed rotation augmentation that introduces controlled variations in skeleton orientation, simulating different poses and viewpoints. By doing so, we strengthen the

dataset's ability to generalize across a wider range of real-world scenarios, improving the model's performance. We also use a balanced data loader during training to ensure that each batch of data contains samples from all classes in equal proportions. This is important to avoid overfitting to majority classes.

*Pre-processing* To ensure consistent origin points across all sequences, we employ a sequence-based normalization technique. In particular, we subtract the spine coordinates of the first frame from each skeleton in the sequence. This enhances the model performance since it standardizes the starting position across all sequences, thereby mitigating potential biases introduced by variations in initial skeleton positions. By relying on such a uniformly set reference frame, the model can more accurately learn the underlying patterns and spatial dependencies of the data without being influenced by irrelevant positional discrepancies. We ultimately excluded the joints of the hands and feet as they were deemed irrelevant and introduced additional noise into the data. It is noteworthy that our comparisons are against state-of-the-art (SOTA) methods employing 25 joints to ensure a fair assessment. In the ablation studies, we presented the findings of the 17-joint configuration to showcase its superior efficacy over the 25-joint configuration in the KIMORE dataset [15].

## 4.3. Implementation Details

The proposed model is trained on the skeletal data of KIMORE and UI-PRMD datasets after data augmentation and pre-processing as explained in section 4.2. The normalized 3D joint positions of the skeletons are used as input to the Spatial GCN. Following [29], the network is trained on a 0.8/0.2 train/test. We utilized the Adam optimizer, known for its efficiency and effectiveness in training deep learning models, with an initial learning rate set to 0.0001. Additionally, we set the batch size to 16, balancing computational efficiency and the stability of gradient updates. A patience value of 500 epochs is set to monitor the validation loss.

An extensive grid search was carried out for selecting hyper-parameters of GCNs (). In particular, we set the number of layers to  $k=2$  for both the spatial and the temporal GCN and use the mean function to aggregate information from adjacent joints at each layer in the two GCNs. We set the number of frames  $M$  in spatial, temporal and our classifiers-based GCN to 100.

We performed all experiments using the PyTorch framework on a machine with an Intel i7 4.20 GHz processor and Tesla T4 graphic card. In the context of human movement quality assessment, the Mean Absolute Deviation ( $MAD$ ) is ordinarily used to measure the difference between the ground truth movement quality scores and the predicted ones:

$$MAD = \frac{1}{b} \sum_{i=1}^b \|y - \bar{y}\| \quad (6)$$

where  $b$  is the sample size and  $y$  and  $\bar{y}$  are the ground-truth and predicted quality score values, respectively.  $MAD$  is a



**Table 1**Results of ten exercises on the UI-PRMD dataset using the evaluation metric *MAD* (bold typeface shows best performances)

Exercise	<i>STGCN-Seq</i>	Deb et al [12]	D-STGCN[42]	Song et al [13]	Zhang et al [27]	Liao et al [9]	Li et al [28]	Shahroudi et al [48]	Du et al [29]
1	<b>0.006</b>	0.009	0.011	0.011	0.022	0.011	0.011	0.018	0.030
2	0.008	<b>0.006</b>	0.009	0.006	0.008	0.028	0.029	0.044	0.077
3	<b>0.009</b>	0.013	0.013	0.010	0.016	0.039	0.056	0.081	0.137
4	<b>0.006</b>	0.006	0.009	0.014	0.016	0.012	0.014	0.024	0.036
5	<b>0.003</b>	0.008	0.009	0.013	0.008	0.019	0.017	0.032	0.064
6	<b>0.004</b>	0.006	0.013	0.009	0.008	0.018	0.019	0.034	0.047
7	<b>0.009</b>	0.011	0.022	0.017	0.021	0.038	0.027	0.049	0.193
8	<b>0.013</b>	0.016	0.020	0.017	0.025	0.023	0.025	0.051	0.073
9	<b>0.006</b>	0.008	0.013	0.008	0.027	0.023	0.027	0.043	0.065
10	0.028	0.031	<b>0.014</b>	0.038	0.066	0.042	0.047	0.077	0.160
Avg	<b>0.009 (-22%)</b>	0.011	0.013	0.014	0.021	0.025	0.027	0.045	0.088

simple and effective measure of model performance where lower values indicate better performance.

In addition to the *MAD* metric, we report Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), which are commonly used in the field of APRE to assess the accuracy of predictive models.

RMSE, measures the average magnitude of the errors between predicted scores and the ground truth. In particular, it is calculated by taking the square root of the average of the squared differences between predicted and actual values. It provides a single measure of the magnitude of prediction errors, with lower values indicating better accuracy. It is sensitive to large errors due to the squaring operation, making it particularly useful for identifying outliers or extreme deviations in predictions. Mathematically, it can be expressed as:

$$RMSE = \sqrt{\frac{1}{b} \sum_{i=1}^b (y - \bar{y})^2} \quad (7)$$

MAPE, on the other hand, measures the average absolute percentage difference between predicted and actual values. Mathematically, it can be expressed as:

$$MAPE = \frac{1}{b} \sum_{i=1}^b \left\| \frac{y - \bar{y}}{y} \right\| \times 100 \quad (8)$$

## 5. Results

In this section, we comprehensively evaluate our findings by juxtaposing them against state-of-the-art methodologies on the two referenced datasets [15, 16]. Additionally, we offer a qualitative analysis that provides further insights into the efficacy of our approach, we conduct several ablation studies to meticulously validate our methodology and finally,

we present the results of our classifier validation, affirming the reliability and accuracy of our models.

### 5.1. Comparison with state-of-the-art approaches

To ensure a fair assessment of our GCN-based model against state-of-the-art deep learning techniques [42, 12, 13, 14, 9, 28, 48, 29, 11], our reported results are without using the classification stage, which involves training the **VC** on a filtered dataset. Our approach, denoted as *STGCN-Seq*, underwent validation using the original datasets employed in the aforementioned studies. We strictly adhered to the same evaluation criteria and training-test partitioning as specified in [9, 12, 42]. In particular, we report the results of a 10-run evaluation to assess the performance of our model. We conduct both training and testing ten times to ensure the reliability of our results. For each run, we record performance metrics such as *MAD*, RMSE, and MAPE. After completing the ten runs, the values of these metrics are averaged to provide a comprehensive measure of the model's performance. The standard deviation across different runs is 0.0001 for the UI-PRMD dataset and 0.002 for the Kimore dataset. These small error bands highlight the low variability and high robustness of the performance of the model. We note that the metric scores provided in our comparison are directly sourced from the original papers, ensuring accuracy and reliability in our reporting. Additionally, the percentage of improvement is calculated relative to the second-best method in the comparison.

Initially, we present our findings based on the analysis conducted on the ten exercises comprising the UI-PRMD dataset. Subsequently, we provide detailed results for each of the five exercises contained within the KIMORE dataset. As illustrated in Table 1 and Table 2, our proposed model exhibits superior performance across multiple evaluation metrics, including *MAD*, RMSE, and MAPE. Notably, we

**Table 2**

Results of five exercises on the KIMORE dataset (bold typeface shows best performances)

Metric	Exercise	STGCN-Seq	D-STGCNT [42]	Deb et al [12]	Song et al [13]	Zhang et al [14]	Liao et al [9]	Yan et al [11]	Li et al [28]	Du et al [29]
MAD	1	<b>0.543</b>	0.641	0.799	0.977	1.757	1.141	0.889	1.378	1.271
	2	<b>0.511</b>	0.753	0.774	1.282	3.139	1.528	2.096	1.877	2.199
	3	0.213	<b>0.210</b>	0.369	1.105	1.737	0.845	0.604	1.452	1.123
	4	<b>0.204</b>	0.206	0.347	0.415	1.202	0.468	0.842	0.675	0.880
	5	0.488	<b>0.399</b>	0.621	1.536	1.853	0.847	1.218	1.662	1.864
	Avg	<b>0.391 (-11%)</b>	0.441	0.582	1.063	1.937	0.965	1.129	1.408	1.467
RMSE	1	<b>1.492</b>	2.020	2.024	2.165	2.916	2.534	2.017	2.344	2.440
	2	<b>1.124</b>	1.468	2.120	3.345	4.140	3.738	3.262	2.823	4.297
	3	<b>0.337</b>	0.487	0.556	1.929	2.615	1.561	0.799	2.004	1.925
	4	<b>0.218</b>	0.527	0.644	2.018	1.836	0.792	1.331	1.078	1.676
	5	<b>0.724</b>	0.735	1.181	3.198	2.916	1.914	1.951	2.575	3.158
	Avg	<b>0.779 (-25%)</b>	1.047	1.305	2.531	2.884	2.108	1.872	2.164	2.699
MAPE	1	<b>1.362</b>	1.623	1.926	2.605	5.054	2.589	2.339	3.491	3.228
	2	<b>0.766</b>	0.974	1.272	3.296	10.436	3.976	6.136	5.298	6.001
	3	0.620	<b>0.613</b>	0.728	2.968	5.774	2.023	1.727	4.188	3.421
	4	<b>0.514</b>	0.541	0.824	2.152	3.901	2.333	2.325	1.976	2.584
	5	1.412	<b>1.217</b>	1.591	4.959	6.531	2.312	3.802	5.752	5.620
	Avg	<b>0.934 (-6%)</b>	0.993	1.268	3.196	6.339	2.647	3.266	4.141	4.170

achieve the lowest average scores on both the UI-PRMD and KIMORE datasets, with particularly significant improvements observed on the KIMORE dataset. The superior performance can be attributed to several factors. Firstly, the KIMORE dataset encompasses a wider array of complex exercises involving both healthy and unhealthy subjects, providing a more challenging and diverse training environment. Additionally, the data collected from the Kinect v2 sensor introduces noise and variability, contrasting with the more precise poses obtained using Vicon in the UI-PRMD dataset. This variability and complexity in the KIMORE dataset pose a greater challenge for developing generalizable models. In contrast, the controlled and straightforward nature of the UI-PRMD dataset allows for various methods to demonstrate relatively comparable performance levels.

To further demonstrate the effectiveness of our method, we visualize the ground truth and predicted movement quality scores for all the test sequences in the KIMORE dataset as shown in Figure 4. As can be observed, the predictions of our model were very close to the clinicians' assessments, which is a strong indicator that our method can capture the subtle nuances of human movement quality (ideal performance is drawn as the diagonal blue line). The commendable results achieved by our method can be attributed to several factors. Firstly, the utilization of per-action adjacency matrices enhances the accuracy of quality score estimation, indicating the efficacy of this approach in capturing the nuanced relationships between joints for each specific exercise. Furthermore, our findings underscore the effectiveness of sequential learning of spatial and temporal features through GCNs.

This methodology demonstrates its potential in achieving good performance.

Finally, our analysis indicates that the model performance varies between healthy and pathological subjects. Specifically, we observed that the model generally achieves higher accuracy in estimating quality scores for sequences collected from healthy subjects (MAD = 0.379) compared to those from pathological subjects (MAD = 0.422). This discrepancy is likely due to the greater variability and complexity in the movement patterns associated with pathological conditions, which can challenge the ability of the model to generalize effectively.

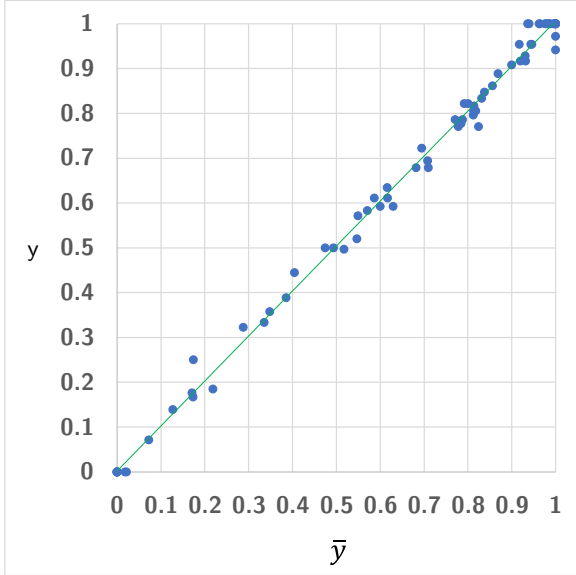
## 5.2. Computational Cost

Our proposed method undergoes rigorous testing, evaluating both computational efficiency and accuracy in quality score estimation on Ex5 of the KIMORE dataset. Testing on a single Tesla T4 GPU with 16 GB of RAM reveals that our model processes a video in 8.3 milliseconds on average, demonstrating its efficiency and capacity for real-time performance. With a modest parameter count of 123K, our model stands out as lightweight compared to contemporary approaches, notably in contrast to Deb et al. [12], which boasts 772K parameters and ranks as the second-best performer in UI-PRMD and third-best in KIMORE. This efficiency is attributed to our approach's avoidance of LSTMs for processing entire 3D skeleton input sequences and eschewing attention mechanisms, which typically demand substantial computational resources. The reported execution times position our method as particularly suitable

**Table 3**

Computational cost for KIMORE dataset.

Method	Phase	# of videos	Score estimation time	# of parameters	Classification time
Deb et al. [12]	Train	373	27h	772K	-
STGCN-Seq	Train	373	58min	123K	50 min
Deb et al. [12]	Test	100	6s	772K	-
STGCN-Seq	Test	100	1s	123K	0.8s

**Fig. 4:** Comparison of the prediction of the proposed approach  $\bar{y}$  against the clinical assessment  $y$  for the KIMORE dataset.

for rehabilitation applications, where timely processing is paramount.

Furthermore, our proposed classification stage requires less time compared to the quality score estimation task. Despite this, it has a crucial impact as it alerts the user (clinician) if the selected sequence does not include a complete or properly executed movement, which is essential for clinical and diagnostic applications.

### 5.3. Feedback and effect of joints in different rehabilitation exercises

In support of our proposal in learning separate adjacency matrices per exercise type (cf. Section 3.2), we provide illustrations of the learnable adjacency matrices, which effectively highlight the significance of joint connections in each specific exercise (see Figure 5).

For instance, examining the adjacency matrix for the squat exercise reveals a pronounced emphasis on the interconnections between the base of the spine and the knees, highlighting their pivotal roles in executing the exercise accurately. Similarly, the adjacency matrix for the pelvis rotation exercise highlights the heightened significance of the spine and the hips, indicating their crucial roles compared to the knees in this specific exercise. Moreover, the findings reveal the model's ability to discern and learn connections

among joints that are physically distant from each other, exemplified by its comprehension of the relationship between the left knee and the right knee in the squatting exercise.

Additionally, Figure 5) presents the Standard Deviation (STD) matrix for all the KIMORE dataset to highlight the effectiveness of adopting a per-action adjacency matrix over a global one for the entire dataset. The non-negligible differences in cell values within the STD matrix suggest that it is more informative to capture the subtle variations unique to each action as opposed to having a single global adjacency matrix, which reinforces the effectiveness of our approach in accurately modeling intricate movement patterns.

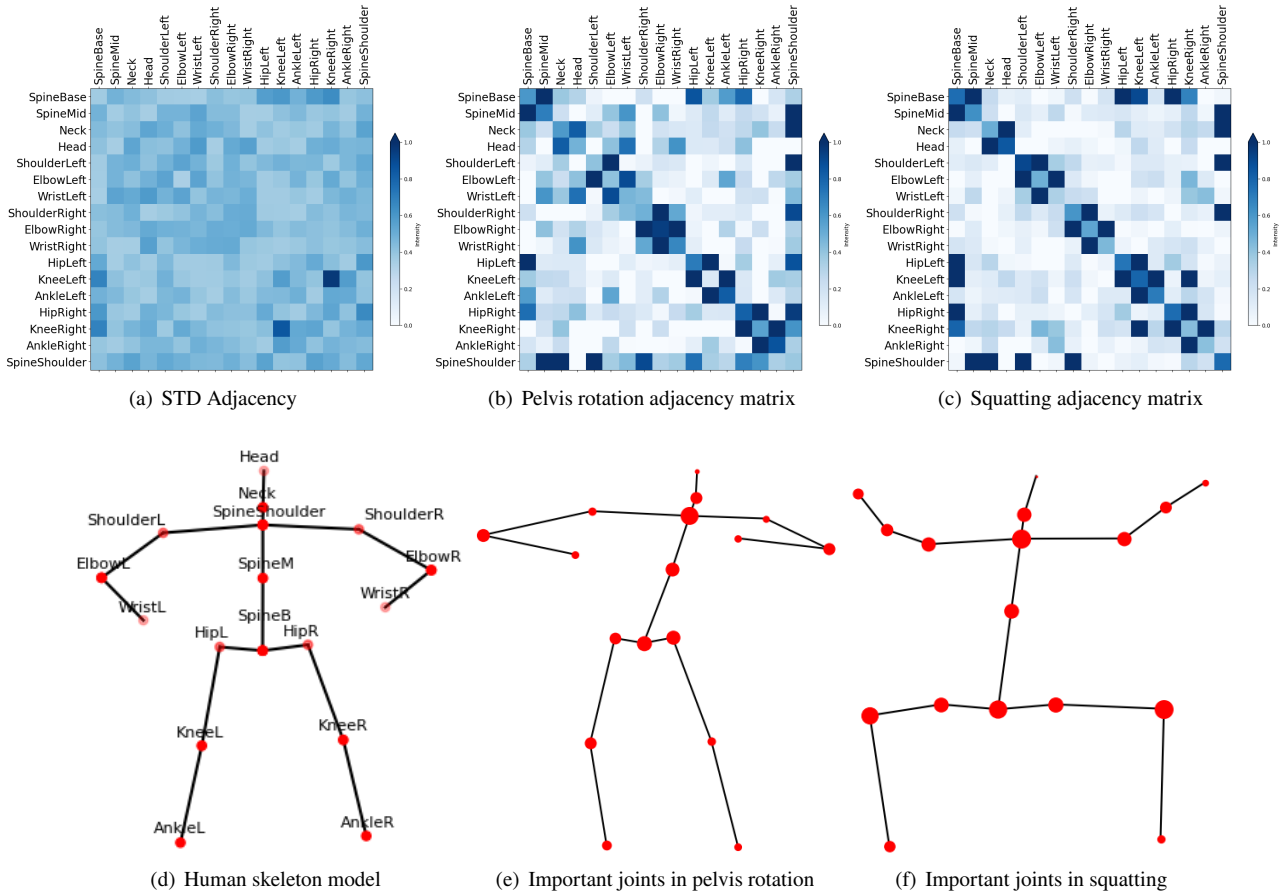
### 5.4. Ablation studies

We conducted several ablation studies in the KIMORE dataset to examine the specific contributions of individual components within our STGCN-Seq model. This decision was pivotal, as the KIMORE dataset encompasses data from both healthy and unhealthy subjects, providing a more comprehensive assessment compared to the UI-PRMD dataset, which is restricted to healthy participants.

In the *B* experiment, we trained a model that learned a global adjacency matrix for all the exercises. Quantitative results in Table 4 shows that learning a global adjacency matrix for the entire dataset exercises results in inferior performance compared to our per-action adjacency matrix *STGCN-Seq*, which further outperforms the model that uses a skeleton-based adjacency matrix (see experiment *C*). In this case, spatial features are learned from only physically connected joints as done in [11].

In the *D* experiment, we investigated the effectiveness of our learnable temporal adjacency matrix. To do so, we trained a model with an adjacency matrix that linked each frame with its immediate next and previous frame. Our results showed that the learnable adjacency matrix significantly improved the performance of our full model. To further prove the effectiveness of our model, we compare it against the *E* experiment that uses an LSTM instead of our temporal GCN. In particular, we implemented two LSTM layers, each comprising 128 hidden units. The Rectified Linear Unit (ReLU) activation function was utilized to introduce non-linearity and facilitate the model's learning process. To prevent overfitting, a dropout rate of 0.1 was applied, randomly deactivating a portion of the neurons during training. For optimization, we employed the Adam optimizer, initializing it with a learning rate of 0.0001. Additionally, the sequence length was set to 100, aligning with the input size of the temporal GCNs.

## Spatiotemporal sequential GCNs



**Fig. 5:** The visualization illustrates the standard deviation matrix for all per-action learnable adjacency matrices within the KIMORE dataset along with the human topology model. Additionally, we provide examples of the pelvis rotation and squatting exercises, displaying their respective learnable adjacency matrices and emphasizing the key joints for each exercise. The size of the circles reflects the significance of the joints in their respective exercises.

**Table 4**

Ablation studies of the proposed approach on the KIMORE dataset using the *MAD* metric.

Method	Per-action	Adjacency					Data			MAD
		Global	Skeleton	Temporal	Prev-next	GCN-LSTM	Augmentation	Normalization	17-joint	
<i>STGCN-Seq</i>	✓			✓			✓	✓		0.391
<i>B</i>		✓		✓			✓	✓		0.704
<i>C</i>			✓	✓			✓	✓		0.548
<i>D</i>	✓				✓		✓	✓		0.586
<i>E</i>	✓					✓	✓	✓		0.553
<i>F</i>	✓			✓				✓		0.591
<i>G</i>	✓			✓			✓			0.454
<i>H</i>	✓			✓			✓		✓	<b>0.379</b>

Table 4 shows that our approach attains better performances using only GCNs, which are more efficient.

In the *F* and *G* experiments, we conducted a comparative analysis of our model’s performance with and without data augmentation and preprocessing. Our findings underscored the critical role of data augmentation in enhancing the model’s overall performance, highlighting its significance in achieving improved results in the *MAD* metric. Additionally,

we observed that the used preprocessing technique helps to get more accurate quality scores.

In the final experiment *H*, our model *STGCN-Seq* was deployed after omitting the joints associated with the hands and feet. Specifically, we utilized 17 joints instead of the 25 joints typically included in the Kinect-based model. Remarkably, we observed a notable improvement in performance on the KIMORE dataset subsequent to this adjustment. This

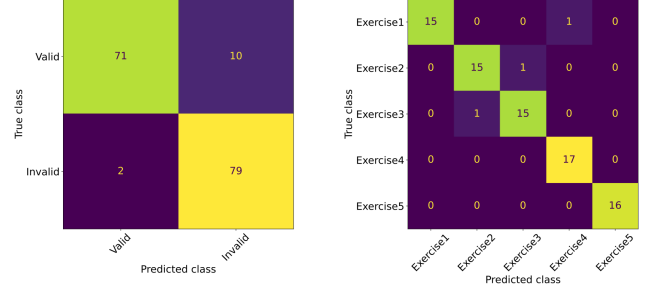
**Table 5**Ablation studies of the used hyper-parameters on the KIMORE dataset using the *MAD* metric.

Method	# of GCN layers			Aggregation function in GCN			50	# of frames			MAD
	1	2	3	mean	max	sum		100	200		
<i>STGCN-Seq</i>		✓		✓				✓			<b>0.391</b>
<i>I</i>	✓			✓				✓			0.677
<i>J</i>			✓	✓				✓			0.421
<i>K</i>		✓			✓			✓			0.412
<i>L</i>		✓				✓		✓			0.399
<i>M</i>		✓		✓			✓				0.430
<i>N</i>		✓		✓					✓		0.473

enhancement can be ascribed to the recognition that these omitted joints hold less relevance for the actions within this dataset. Moreover, their exclusion helped mitigate noise and pose tracking complexities, especially concerning the Kinect V2 sensor's performance.

In the second part of our ablation studies, we conducted a comprehensive evaluation to determine the optimal hyper-parameters for our GCNs (Table 5). We tested various configurations, including the number of layers and aggregation functions. Additionally, we selected the optimal number of frames *M* to resize the input sequence.

- **Number of layers:** we experimented with using 1, 2, and 3 layers (baseline *I*, baseline *STGCN-seq* experiment, and baseline *J*) in both the spatial and temporal GCN components. The purpose of these tests was to determine the optimal depth for capturing spatial and temporal dependencies in the data. (Table 5 indicates that a two-layer configuration achieves the best performance, likely due to its balance between model complexity and the ability to capture essential features without overfitting.
- **Aggregation function:** we explored different strategies for aggregating information from adjacent joints within each layer of the GCNs. We compared the use of maximum (baseline *K*), summation (baseline *L*), and average (mean) functions to integrate the features from neighboring nodes. The results in Table 5 demonstrated that using the mean function in baseline *STGCN-seq* consistently provided better performance than the other methods. This may be attributed to the mean function's ability to smooth out noise and maintain a more stable representation of the joint information, which is crucial for accurate movement analysis and prediction.
- **Number of frames:** we also investigated the impact of selecting different numbers of frames *M* on the performance of our model. We observed that using 100 frames provided a balanced representation of the input sequence, capturing sufficient temporal information without overwhelming the model with excessive data (Table 5). In contrast, using only 50 frames (baseline *M*) resulted in a loss of crucial temporal details,

**Fig. 6:** Confusion matrix of our exercise classifier and validity classifier on the KIMORE dataset.

while 200 frames (baseline *N*) introduced unnecessary complexity and noise, ultimately degrading the model's performance.

### 5.5. Validation of the classification module

Beyond assessing the quality score estimation, we offer a thorough evaluation of the action classification task's performance. The accuracy metric yielded a score of 0.96 for the exercise classifier and 0.93 for the validity classifier. Furthermore, we provide the confusion matrices of both classifiers in Figure 6, enabling a detailed examination of the classification performance across different action categories.

We also assessed the impact of our proposed classification stage on the quality score regression performance. The MAD results, with a value of 0.348, demonstrate a significant enhancement, indicating an 11% improvement compared to the model without the classification stage (MAD=0.391). This highlights the effectiveness of the validity classifier, which not only constrains the model's output to valid sequences but also strengthens the robustness and accuracy of our exercise performance assessment.

## 6. Conclusions and future work

In this work, we adopt GCNs to learn spatial and temporal features for APRE. Our approach attained performance superior to the state-of-the-art methods since it learns the implicit connections between joints using a single model. This allows our model to capture the global structure of the human body and the relationships between different body parts, which is essential for understanding and analyzing

complex movements. Furthermore, our study demonstrates that integrating a classification stage before evaluating the quality score prompts the model to generate quality scores exclusively for valid sequences. This refinement improves the robustness and accuracy of our model's predictions.

As part of our future work, we will investigate the use of different types of adjacency matrices to further improve the performance and generalization capacity of our model. For example, we could explore the use of weighted adjacency matrices during the same exercise. At each moment, we could assign different weights to different connections between joints based on their importance. We could also explore the use of adjacency matrices that are specific to different body parts, which would allow our model to better capture the unique joint relationships in each body part.

Additionally, the findings of this study underscore the critical need for comprehensive rehabilitation exercise datasets. These datasets should feature balanced class distributions across diverse health conditions and incorporate controlled conditions during exercise execution. Such an approach ensures that the dataset reflects the full spectrum of scenarios encountered in real-world rehabilitation settings, thereby enabling more robust and generalizable model training. Moreover, the dataset will include various exercises, allowing us to further explore and demonstrate the versatility and effectiveness of our learnable adjacency matrices across a broader range of movements.

## 7. Acknowledgement

This research was funded by the Région Bretagne and the Conseil départemental du Finistère through the ECFvisuL project, as well as the Institut Carnot Télécom & Société numérique through the FCEval project.

This research benefited from the guidance of Dr. Brice Loddé, Dr. Pierre Balla, and Dr. Thomas Le Rhun, occupational health experts at CHRU Brest, as well as Mrs. Anna Thepaut, a physiotherapist at Pôle Kiné Plouzané.

## License

In accordance with our funding institution's rules regarding open access to results of publicly funded scientific research, the current and all subsequent versions of this article will be published under CC-BY 4.0 license.

## References

- [1] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, D. Tao, Tridet: Temporal action detection with relative boundary modeling, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 18857–18866.
- [2] Y. Mao, J. Deng, W. Zhou, Y. Fang, W. Ouyang, H. Li, Masked motion predictors are strong 3d action representation learners, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10181–10191.
- [3] W. Xin, R. Liu, Y. Liu, Y. Chen, W. Yu, Q. Miao, Transformer for skeleton-based action recognition: A review of recent advances, *Neurocomputing* (2023).
- [4] M. G. Morshed, T. Sultana, A. Alam, Y.-K. Lee, Human action recognition: A taxonomy-based survey, updates, and opportunities, *Sensors* 23 (2023) 2182.
- [5] E. Stawiarska, M. Stawiarski, Assessment of patient treatment and rehabilitation processes using electromyography signals and selected industry 4.0 solutions, *International Journal of Environmental Research and Public Health* 20 (2023) 3754.
- [6] M. H. Lee, D. P. Siewiorek, A. Smaligic, A. Bernardino, S. B. i. Badia, Learning to assess the quality of stroke rehabilitation exercises, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 218–228.
- [7] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, *Advances in neural information processing systems* 25 (2012).
- [8] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, J. Schmidhuber, Lstm: A search space odyssey, *IEEE transactions on neural networks and learning systems* 28 (2016) 2222–2232.
- [9] Y. Liao, A. Vakanski, M. Xian, A deep learning framework for assessing physical rehabilitation exercises, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28 (2020) 468–477.
- [10] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, *arXiv preprint arXiv:1609.02907* (2016).
- [11] S. Yan, Y. Xiong, D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.
- [12] S. Deb, M. F. Islam, S. Rahman, S. Rahman, Graph convolutional networks for assessment of physical rehabilitation exercises, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022) 410–419.
- [13] Y.-F. Song, Z. Zhang, C. Shan, L. Wang, Richly activated graph convolutional network for robust skeleton-based action recognition, *IEEE Transactions on Circuits and Systems for Video Technology* 31 (2020) 1915–1925.
- [14] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, D.-S. Chen, A comprehensive survey of vision-based human action recognition methods, *Sensors* 19 (2019) 1005.
- [15] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, A. Monteriu, L. Romeo, F. Verdini, The kimore dataset: Kinematic assessment of movement and clinical scores for remote monitoring of physical rehabilitation, *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27 (2019) 1436–1448.
- [16] A. Vakanski, H.-p. Jun, D. Paul, R. Baker, A data set of human body movements for physical rehabilitation exercises, *Data* 3 (2018) 2.
- [17] S. Sardari, S. Sharifzadeh, A. Daneshkhan, B. Nakisa, S. W. Loke, V. Palade, M. J. Duncan, Artificial intelligence for skeleton-based physical rehabilitation action evaluation: A systematic review, *Computers in Biology and Medicine* (2023) 106835.
- [18] A. Nogales, M. Rodríguez-Aragón, Á. J. García-Tejedor, A systematic review of the application of deep learning techniques in the physiotherapeutic therapy of musculoskeletal pathologies, *Computers in Biology and Medicine* 172 (2024) 108082.
- [19] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, V. Kyriki, A. Monteriu, L. Romeo, F. Verdini, A hidden semi-markov model based approach for rehabilitation exercise assessment, *Journal of biomedical informatics* 78 (2018) 1–11.
- [20] J. F.-S. Lin, M. Karg, D. Kulić, Movement primitive segmentation for human motion modeling: A framework for analysis, *IEEE Transactions on Human-Machine Systems* 46 (2016) 325–339.
- [21] A. Vakanski, J. Ferguson, S. Lee, Mathematical modeling and evaluation of human motions in physical therapy using mixture density neural networks, *Journal of physiotherapy & physical rehabilitation* 1 (2016).
- [22] V. Antoniou, C. H. Davos, E. Kapreli, L. Batalik, D. B. Panagiotakos, G. Pepera, Effectiveness of home-based cardiac rehabilitation, using wearable sensors, as a multicomponent, cutting-edge intervention: a systematic review and meta-analysis, *Journal of clinical medicine* 11 (2022) 3772.

- [23] C. E. Lang, J. Barth, C. L. Holleran, J. D. Konrad, M. D. Bland, Implementation of wearable sensing technology for movement: pushing forward into the routine physical rehabilitation care field, *Sensors* 20 (2020) 5744.
- [24] K. Réby, I. Dulau, G. Dubrasquet, M. B. Aimar, Graph transformer for physical rehabilitation evaluation, in: 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), IEEE, 2023, pp. 1–8.
- [25] M. Khodatars, A. Shoeibi, D. Sadeghi, N. Ghaasemi, M. Jafari, P. Moridian, A. Khadem, R. Alizadehsani, A. Zare, Y. Kong, et al., Deep learning for neuroimaging-based diagnosis and rehabilitation of autism spectrum disorder: a review, *Computers in Biology and Medicine* 139 (2021) 104949.
- [26] T. Hamaguchi, T. Saito, M. Suzuki, T. Ishioka, Y. Tomisawa, N. Nakaya, M. Abo, Support vector machine-based classifier for the assessment of finger movement of stroke patients undergoing rehabilitation, *Journal of Medical and Biological Engineering* 40 (2020) 91–100.
- [27] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 1112–1121.
- [28] C. Li, Q. Zhong, D. Xie, S. Pu, Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation, *arXiv preprint arXiv:1804.06055* (2018).
- [29] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1110–1118.
- [30] A. Kanade, M. Sharma, M. Muniyandi, A robust and scalable attention guided deep learning framework for movement quality assessment, *arXiv preprint arXiv:2204.07840* (2022).
- [31] J. Liu, X. Wang, C. Wang, Y. Gao, M. Liu, Temporal decoupling graph convolutional network for skeleton-based gesture recognition, *IEEE Transactions on Multimedia* (2023).
- [32] A. S. M. Miah, M. A. M. Hasan, J. Shin, Dynamic hand gesture recognition using multi-branch attention based graph and general deep learning model, *IEEE Access* 11 (2023) 4703–4716.
- [33] W. Yang, J. Zhang, J. Cai, Z. Xu, Hybridnet: Integrating gcn and cnn for skeleton-based action recognition, *Applied Intelligence* 53 (2023) 574–585.
- [34] K. Hu, J. Jin, C. Shen, M. Xia, L. Weng, Attentional weighting strategy-based dynamic gcn for skeleton-based action recognition, *Multimedia Systems* 29 (2023) 1941–1954.
- [35] J. Lee, M. Lee, D. Lee, S. Lee, Hierarchically decomposed graph convolutional networks for skeleton-based action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10444–10453.
- [36] R. Hou, Z. Wang, R. Ren, Y. Cao, Z. Wang, Multi-channel network: Constructing efficient gcn baselines for skeleton-based action recognition, *Computers & Graphics* 110 (2023) 111–117.
- [37] K. Zhou, Y. Ma, H. P. Shum, X. Liang, Hierarchical graph convolutional networks for action quality assessment, *IEEE Transactions on Circuits and Systems for Video Technology* (2023).
- [38] Z. Shan, Q. Yang, R. Ye, Y. Zhang, Y. Xu, X. Xu, S. Liu, Gpa-net: No-reference point cloud quality assessment with multi-task graph convolutional network, *IEEE Transactions on Visualization and Computer Graphics* (2023).
- [39] Q. Lei, H. Li, H. Zhang, J. Du, S. Gao, Multi-skeleton structures graph convolutional network for action quality assessment in long videos, *Applied Intelligence* 53 (2023) 21692–21705.
- [40] S. H. Chowdhury, M. Al Amin, A. M. Rahman, M. A. Amin, A. A. Ali, Assessment of rehabilitation exercises from depth sensor data, in: 2021 24th International Conference on Computer and Information Technology (ICCCIT), IEEE, 2021, pp. 1–7.
- [41] M. Chen, Y. Chen, Y. Xu, Q. An, W. Min, Population flow based spatial-temporal eigenvector filtering modeling for exploring effects of health risk factors on covid-19, *Sustainable Cities and Society* 87 (2022) 104256.
- [42] Y. Mouchid, R. Slama, D-stgcnt: A dense spatio-temporal graph conv-gru network based on transformer for assessment of patient physical rehabilitation, *Computers in Biology and Medicine* 165 (2023) 107420.
- [43] K. Shiraki, T. Hirakawa, T. Yamashita, H. Fujiyoshi, Acquisition of optimal connection patterns for skeleton-based action recognition with graph convolutional networks., in: VISIGRAPP (5: VISAPP), 2020, pp. 302–309.
- [44] J. Xie, Q. Miao, R. Liu, W. Xin, L. Tang, S. Zhong, X. Gao, Attention adjacency matrix based graph convolutional networks for skeleton-based action recognition, *Neurocomputing* 440 (2021) 230–239.
- [45] R. Yang, J. Niu, Y. Xu, Y. Wang, L. Qiu, Action recognition based on gcn with adjacency matrix generation module and time domain attention mechanism, *Symmetry* 15 (2023) 1954.
- [46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [47] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1010–1019.
- [48] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang, Ntu rgb+ d: A large scale dataset for 3d human activity analysis, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1010–1019.